ACENET

Microcredential in Advanced Computing

ISP Report *Template*

**Project title:** Predicting Hospital Stay Durations Using Machine Learning

**Participant name:** Maria Hennebury

**Date:**

**Abstract:** This project aims to predict hospital stay durations for patients in Newfoundland and Labrador using machine learning techniques. By analyzing various contributing factors such as patient demographics, morbidity trends, and hospital resources, the goal is to enhance resource allocation and improve healthcare efficiency.

**1. Introduction**

The healthcare system in Newfoundland and Labrador faces challenges including rising morbidity rates and the need for efficient resource management. This project addresses the problem of predicting hospital stay durations to optimize patient care and resource allocation. The research question is: Can machine learning models accurately predict the length of hospital stays based on patient demographics and morbidity data?

**2. Background**

Hospital stay duration is a critical metric in healthcare management, influencing resource allocation, patient care, and hospital efficiency. Understanding the factors that affect the length of stay can lead to better planning and reduced costs. This project explores how various factors such as patient age, gender, and specific diseases influence hospital stay lengths.

The dataset used in this project contains records of hospital stays, with the key variable being 'Days in hospital (average)'. Summary statistics reveal that:

* **Count:** 13,979 observations
* **Mean:** 7.98 days
* **Standard Deviation:** 2.57 days
* **Minimum:** 3 days
* **25th Percentile:** 6 days
* **Median:** 8 days
* **75th Percentile:** 9 days
* **Maximum:** 64 days

**Interpretation:**

* **Central Tendency:** The mean and median values being close suggest that the distribution of hospital stays is relatively symmetric with a central tendency around 8 days.
* **Dispersion:** The standard deviation indicates moderate variability around the mean, while the range (3 to 64 days) suggests the presence of outliers.
* **Distribution Shape:** The close proximity of the 25th and 75th percentiles to the median implies a relatively symmetric distribution, though the high maximum value indicates some extreme cases.

**3. Analysis**

Describe the approach you took to complete your project. Explain where you found your dataset, and explain the dataset. Describe and justify the data preparation process. Describe your choice of analysis method (for example if you used ML, which model did you use and why). Describe how you used HPC for your project.

**4. Results**

The Random Forest model was evaluated using the following metrics:

* **Mean Squared Error (MSE):** 0.627
* **R-squared (R²):** 0.770
* **Mean Absolute Error (MAE):** 0.546

**Findings:**

* The model explained approximately 77% of the variance in hospital stay durations.
* The MSE and MAE indicate a reasonable level of accuracy, though there is room for improvement.

**Plots/Graphs:**

* Include visualizations of the distribution of hospital stays, comparisons of predicted vs. actual values, and feature importance plots.

**5. Discussion**

Analyze your results and their implications. Discuss any limitations or challenges you encountered during the project.

**Results Analysis:**

The model's performance suggests a good fit, with significant variance explained by the features. The presence of outliers, especially those with extremely high values (e.g., 64 days), impacts model accuracy and warrants further investigation.

**Challenges:**

* Handling extreme outliers proved challenging. Their impact on model performance highlighted the need for robust outlier handling strategies.

**Limitations:**

* The dataset may have inherent biases or missing features that could affect the generalizability of the model.

**Future Research:**

* Further investigation into outlier handling and feature engineering could enhance model performance. Additionally, exploring other machine learning algorithms and incorporating external datasets could provide a more comprehensive analysis.

**Conclusion**

Briefly summarize your main findings and conclusions. Restate the significance of your project. Suggest directions for future research based on what you found.

The project demonstrated that machine learning models, specifically Random Forest Regressors, can predict hospital stay durations with considerable accuracy. Understanding the variability in hospital stays, particularly with extreme outliers, is crucial for improving model performance and healthcare resource allocation. Future research should focus on refining outlier handling methods and exploring additional features to further enhance predictive capabilities.

**References**

Include references to the sources for your data, background information, and previous studies used for your project.

**Supplementary Materials**

<https://opendata.gov.nl.ca/public/opendata/page/?page-id=datasetdetails&id=69>

<https://github.com/MariaHennebury/ISP.git>